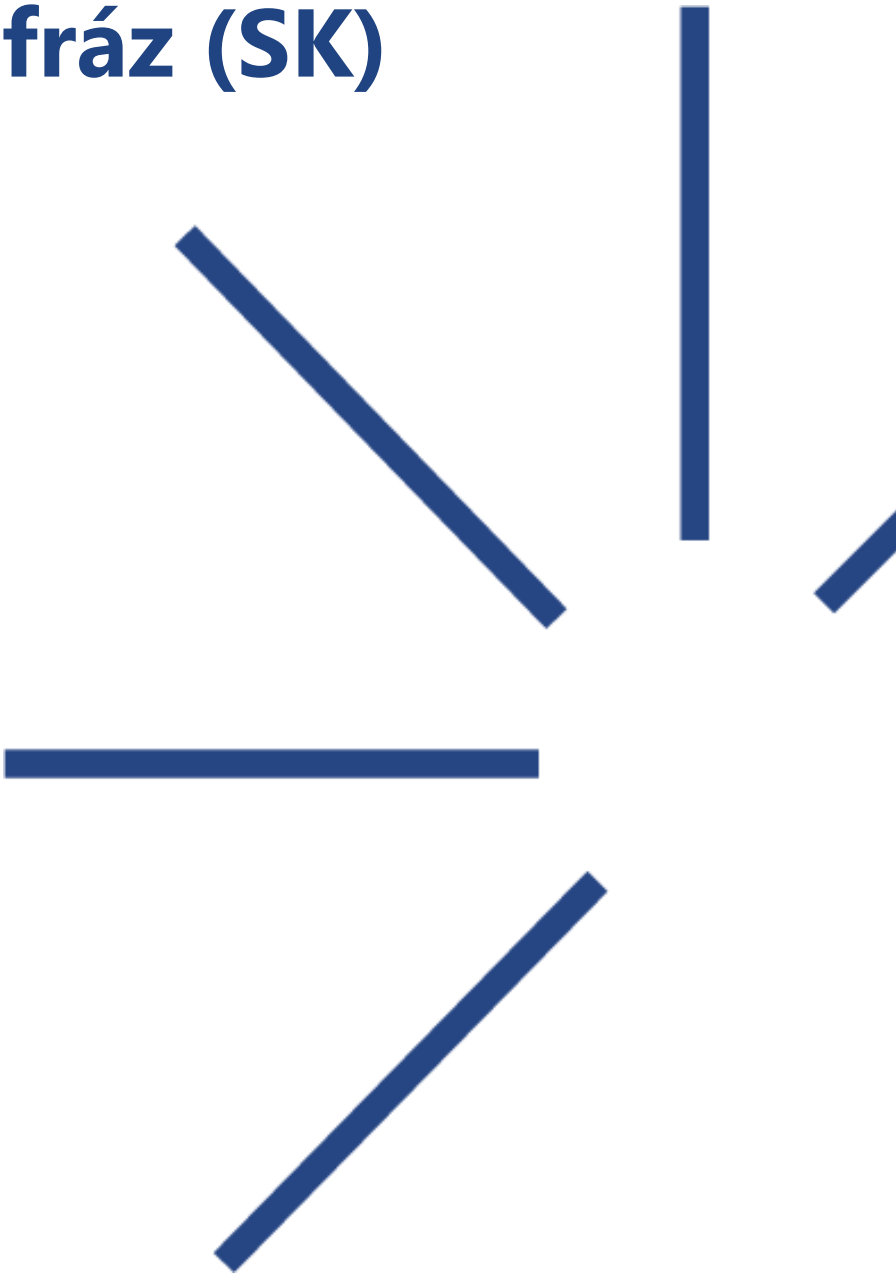
D07 – Metóda na extrakciu procesne súvisiacich fráz (SK)



Projekt financovaný Európskou úniou Next Generation EU prostredníctvom Plánu obnovy a odolnosti SR v rámci projektu č. 09I05-03-V02-00049.


Financované
Európskou úniou
NextGenerationEU

PLÁN OBNOVY

ÚRAD PODPREDSEDU VLÁDY
SLOVENSKEJ REPUBLIKY
PRE PLÁN OBNOVY
A ZNALOSTNÚ EKONOMIKU

VAIA
VÝSKUMNÁ
A INOVAČNÁ
AUTORITA

VÝSKUMNÁ
AGENTÚRA

ÚVOD

Projekt „**Automatizácia analýzy právnych textov na základe strojového učenia**“ (ďalej len „**ALTAML**“) predstavuje kľúčovú iniciatívu zameranú na integráciu inovatívnych prístupov v oblasti spracovania údajov a následnej analýzy, so zameraní na oblasť práva, ktorá zahŕňa aj právo informačných a komunikačných technológií. Cieľom tohto projektu je vyvinúť a overiť účinné metódy automatizovanej analýzy právnych textov pomocou techník strojového učenia. To zahŕňa najmä vývoj nástrojov, ktoré uľahčujú spracovanie a analýzu veľkých objemov údajov vo forme rôznych právnych dokumentov, ako aj extrakciu relevantných informácií (atribútov) z takýchto dokumentov vrátane identifikácie kľúčových pojmov, odkazov na právne predpisy a ďalších atribútov.

Cieľom projektu ALTAML je teda prispieť k efektívnejšiemu prístupu k právnym informáciám a k zrýchleniu právnych procesov, čím sa zabezpečí vyššia miera právnej istoty a zlepší sa dostupnosť právnych textov, výsledkov ich analýzy a relevantných právnych informácií pre odbornú aj širokú verejnosť.

V rámci pracovného balíka KPB2 bol vypracovaný výstup Metóda na extrakciu procesne súvisiacich fráz. Navrhnutý systém predstavuje hybridný extraktor procesných právnych udalostí, ktorý kombinuje regulárne výrazy, sémantické vektorové reprezentácie a veľký jazykový model. Cieľom systému je automatická identifikácia procesných úkonov a výsledkov konaní zo slovenských súdnych rozhodnutí, ako sú zamietnutie odvolania, zrušenie rozhodnutia, zastavenie konania alebo vrátenie veci na ďalšie konanie.

Extrakcia je obmedzená na vopred definovanú ontológiu procesných kategórií, čím je zabezpečená konzistentnosť výstupov a minimalizované riziko nesprávnej interpretácie právneho textu. Kombinácia pravidlového a sémantického prístupu umožňuje efektívne spracovanie aj jazykovo variabilných formulácií typických pre slovenské súdne rozhodnutia.

Systém využíva viacvrstvový proces identifikácie procesných udalostí. V prvej fáze sú aplikované regulárne výrazy nad textom rozhodnutia, ktoré umožňujú presnú identifikáciu explicitných procesných formulácií. V prípade, že pravidlový prístup neposkytne výsledok, systém využíva lokálne nasadený veľký jazykový model na extrakciu kandidátnych procesných fráz. Tieto frázy sú následne sémanticky klasifikované pomocou embeddingov vytvorených jazykovým modelom SlovakBERT doladeným na úlohu sémantickej podobnosti viet.

Výsledkom je štruktúrovaná množina procesných udalostí obsahujúca identifikovanú kategóriu udalosti, extrahovanú procesnú frázu, skóre sémantickej

podobnosti a pôvodnú vetu rozhodnutia, z ktorej bola udalosť extrahovaná. Stručnejší popis systému je uvedený nižšie.

Hybridný extraktor procesných udalostí zo súdnych rozhodnutí

Extraktor procesných udalostí je softvérový modul implementovaný v jazyku Python, navrhnutý na automatickú identifikáciu procesných právnych úkonov zo slovenských súdnych rozhodnutí. Systém kombinuje pravidlové metódy spracovania prirodzeného jazyka, sémantické vektorové reprezentácie a veľký jazykový model (LLM) na dosiahnutie robustných a terminologicky presných výsledkov.

Systém je orientovaný na identifikáciu procesných výsledkov konania, najmä:

- zamietnutie odvolania,
- úspešné odvolanie,
- zastavenie konania,
- zrušenie rozhodnutia,
- vrátenie veci na ďalšie konanie.

Funkcionalita systému

Hlavná funkcionalita extraktora pozostáva z nasledujúcich krokov:

1. Príprava vstupných dát

Systém pracuje so štruktúrovaným korpusom slovenských súdnych rozhodnutí vo formáte JSON. Pre každé rozhodnutie je spracovávaný úplný text rozhodnutia.

Texty sú pred spracovaním normalizované jednoduchými transformačnými pravidlami:

- odstránenie nadbytočných medzier,
- zjednotenie fragmentovaných nadpisov,
- úprava formátovania textu.

Po normalizácii je text rozdelený na jednotlivé vety, ktoré predstavujú základnú jednotku ďalšej analýzy.

1.1. Filtrácia významových viet

Nie všetky vety rozhodnutia obsahujú procesne relevantné informácie. Systém preto vykonáva heuristickú filtráciu viet na základe:

- prítomnosti procesných kľúčových slov,
- minimálnej a maximálnej dĺžky viet,
- výskytu právnej terminológie typickej pre výrokové časti rozhodnutí.

Týmto krokom sa výrazne redukuje počet analyzovaných viet a zvyšuje efektivita spracovania.

1.2. Definícia procesnej ontológie

Procesné udalosti sú reprezentované prostredníctvom vopred definovanej ontológie procesných kategórií. Každá kategória obsahuje množinu reprezentatívnych procesných formulácií používaných ako referenčné prototypy pri sémantickom porovnávaní.

Príklady kategórií:

- odvolanie_zamietnute,
- odvolanie_uspesne,
- konanie_zastavene,
- vec_vratena,
- rozhodnutie_zrusene.

2. Extrakcia procesných udalostí

V prvej fáze systém využíva regulárne výrazy na identifikáciu explicitných procesných formulácií v texte rozhodnutia. Tento prístup umožňuje veľmi presnú detekciu štandardizovaných právnych viet.

Príklady detegovaných formulácií:

- „odvolanie sa zamieta“,
- „napadnutý rozsudok sa zrušuje“,
- „konanie sa zastavuje“.

Každá identifikovaná fráza je okamžite priradená ku konkrétnej procesnej kategórii.
Extrakcia pomocou jazykového modelu

Ak pravidlový prístup neidentifikuje žiadnu procesnú udalosť, systém využije lokálne nasadený veľký jazykový model prostredníctvom nástroja Ollama.

Jazykový model je riadený striktne definovaným promptom v slovenskom jazyku, ktorý zabezpečuje:

- extrakciu iba procesných právnych fráz,
- generovanie stručných výstupov bez vysvetlení,
- oddelenie jednotlivých fráz na samostatné riadky,
- deterministické správanie modelu.

Tento krok umožňuje identifikovať aj menej štandardné alebo jazykovo variabilné formulácie procesných úkonov.

3. Sémantická klasifikácia udalostí

Extrahované procesné frázy sú následne transformované do vektorovej reprezentácie pomocou modelu SlovakBERT doladeného na úlohu sémantickej podobnosti viet.

Pre každú procesnú frázu sa vypočíta kosínová podobnosť medzi embeddingom frázy a embeddingami referenčných formulácií jednotlivých procesných kategórií.

Procesná udalosť je priradená ku kategórii s najvyššou sémantickou podobnosťou, pričom:

- je vypočítané skóre podobnosti,
- je identifikovaný najbližší referenčný prototyp,
- je aplikovaný minimálny prah podobnosti.

Tento prístup umožňuje robustnú klasifikáciu aj pri jazykových variáciách právnych formulácií.

4. Postprocessing a deduplikácia

Po extrakcii a klasifikácii udalostí systém vykonáva odstránenie duplicitných záznamov. Duplicity sú identifikované na základe kombinácie:

- kategórie udalosti,
- normalizovanej textovej reprezentácie frázy.

Výsledné udalosti sú následne zoradené podľa skóre sémantickej podobnosti.

Výstup systému

Výstupom systému je JSON štruktúra obsahujúca:

- identifikátor rozhodnutia,
- počet identifikovaných udalostí,
- zoznam procesných udalostí.

Každá procesná udalosť obsahuje

- kategóriu udalosti,
- extrahovanú frázu,
- skóre podobnosti,
- referenčný prototyp,
- pôvodnú vetu rozhodnutia,
- metódu extrakcie.

Výsledný výstup umožňuje ďalšie analytické spracovanie procesných rozhodnutí a vytváranie štruktúrovaných databáz procesných udalostí.

Požiadavky a technológie

Systém je implementovaný v jazyku Python a využíva nasledujúce technológie:

- Sentence Transformers pre sémantické embeddingy,
- SlovakBERT STS model pre reprezentáciu právnych viet,
- scikit-learn pre výpočet kosínovej podobnosti,
- regulárne výrazy (regex) pre pravidlovú extrakciu,
- lokálny LLM (LLaMA 3.3) nasadený cez Ollama,
- CUDA akceleráciu pre efektívne spracovanie embeddingov.

Odporúča sa:

- RAM minimálne 32 GB,
- GPU s podporou CUDA,
- lokálne nasadenie jazykového modelu pre vysokovýkonné dávkové spracovanie rozsiahlych právnych korpusov.

Zdrojový kód projektu a dátový súbor sú udržiavané v repozitári a sú dostupné na adrese: https://gitlab.science.upjs.sk/altaml/process-related-phrases